AI for a Social World – A Social World for AI

Christian J. Feldbacher-Escamilla

Spring 2021

Introduction

AI is increasingly used for social enhancement.

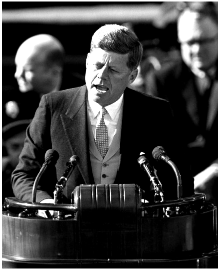
However, is there also a possibility to socially enhance AI?

Our question is not about training AI for a better social integration.

Rather, it is about changing our social structures in favour of AI.

Slogan

"Ask not what your [AI] can do for you, ask what you can do for your [AI]."



Aim for today: Contextualisation and discussion based on a "case study"

Contents

AI for a Social World

The diverse field of AI

The field of AI systematised by the pairs human vs. ideal rationality and reason- vs. action-based (cf. Russell and Norvig 2020, p.2):

	human rationality	ideal rationality
reason-based	thinking humanly	thinking rationally
action-based	acting humanly	acting rationally

Subdisciplines (cf. Russell and Norvig 2020, sect.1.4; and Hauser 2012, sect.3b):

- robotics
- logistics/planning
- game playing
- theorem proving
- natural language processing
- connectionism/neural networks
- knowledge representation
- machine learning

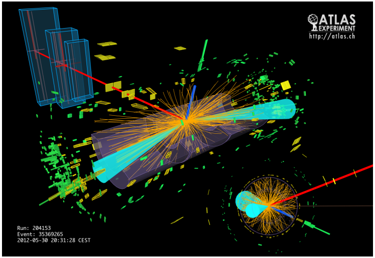
General Applications: Robotics

General Applications: Game Playing

General Applications: Science

CERN's so-called *Higgs boson machine-learning challenge* (CERN 2014):

> *"The goal of the Higgs boson machine-learning challenge is to explore the potential of advanced machine-learning methods to improve the analysis of data produced by the experiment."*
>
> *"Using simulated data with features characterizing events detected by ATLAS, your task is to classify events into 'tau tau decay of a Higgs boson' versus 'background'."*
>
> *"Interested in machine learning? Now is your chance to teach the machines and improve humankind's understanding of the universe."*



Social Impact: Law

Jurisdiction:

> *"Shortly after arrest, a judge has to decide: will the defendant await their legal fate at home? Or must they wait in jail? [...] By law, the judge has to make a prediction: if released, will the defendant return for their court appearance, or will they skip court? And will they potentially commit further crimes?"*

> *"We find that there is considerable room to improve on judges' predictions. [... If we were] using our algorithm's predictions of risk instead of relying on judge intuition, we could reduce crimes committed by released defendants by up to 25% without having to jail any additional people. Or, without increasing the crime rate at all, we could jail up to 42% fewer people." (cf. Kleinberg, Ludwig, and Mullainathan 2016)*

Social Impact: Environment

Rainforest Protection:

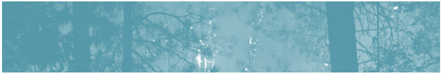
> *"For example, robots with AI capabilities can be used to sort recyclable material from waste. The Rainforest Connection, a Bay Area nonprofit, uses AI tools such as Google's TensorFlow in conservation efforts across the world. Its platform can detect illegal logging in vulnerable forest areas through analysis of audio sensor data. Other applications include using satellite imagery to predict routes and behavior of illegal fishing vessels." (Chui et al. 2018, p.6)*

Social Impact: Crisis, Health, Tax, Etc.

TIME

TECH • WILDFIRES

AI Is Helping Fight Wildfires Before They Start



Stanford | News

Search Stanford news...

Home Find Stories For Journalists Contact

JANUARY 25, 2017

Deep learning algorithm does as well as dermatologists in identifying skin cancer

In hopes of creating better access to medical care, Stanford researchers have trained an algorithm to diagnose skin cancer.

The New York Times

Computer Scientists Wield Artificial Intelligence to Battle Tax Evasion

A Final Example: Communication and Spam

Communication:

> *"Spam fighting: Each day, learning algorithms classify over a billion messages as spam, saving the recipient from having to waste time deleting what, for many users, could comprise 80% or 90% of all messages, if not classified away by algorithms." (Russell and Norvig 2020, p.29)*

Study of AI's Impact

Survey of Chui et al. (2018):

> *"Through an analysis of about 160 AI social impact use cases, we have identified and characterized ten domains where adding AI to the solution mix could have large-scale social impact. [...] Real-life examples show AI already being applied to some degree in about one-third of these use cases, ranging from helping blind people navigate their surroundings to aiding disaster relief efforts."*

Machine learning is now widely used in commercial applications. It's utilisation for solving policy problems is relatively new (cf. Kleinberg, Ludwig, and Mullainathan 2016).

Reversing the Direction

So, there is already a broad and diverse field of impact of AI on society.

But how about our slogan Kennedy style?

What can we do for AI (next to creating it)?

We will have a look at a particular branch of machine learning that has quite a lot to do with spam detection: online machine learning.

The Case of Online Machine Learning

Learning

Several types of *learning* can be distinguished on the basis of the following parameters (see Shalev-Shwartz and Ben-David 2014, sect.1.3):

1. *supervised vs. unsupervised* ... which is about feedback regarding the true outcome/correct classification

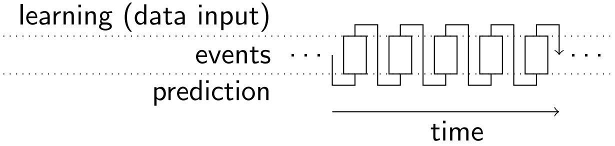
2. *active vs. passive* ... which is about the possibility of interventions

3. *non-adversarial vs. adversarial* ... which is about prediction-related biases in the presentation of samples

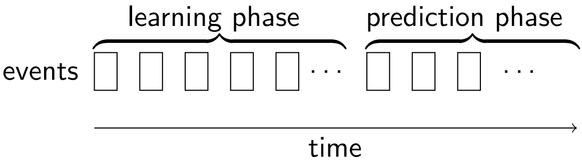
4. *sample based vs. online* ... which is about the learning process, whether it happens in large or small steps

Online vs. Batch Learning

Online Learning:



Sample-Based/Batch Learning:



Learning Paradigms

non-adversarial	supervised	active	sample-based	example
0	0	0	0	Unlucky guessing
0	0	0	1	Unsuccessful ordinary anomaly detection
0	0	1	0	
0	0	1	1	Unsuccessful interactive anomaly detection
0	1	0	0	**Spam detection**
0	1	0	1	Ordinary data mining
0	1	1	0	
0	1	1	1	Anti-realistic ordinary science or student learning "by help" of a hostile instructor in a lab
1	0	0	0	Lucky guessing
1	0	0	1	Successful ordinary anomaly detection
1	0	1	0	
1	0	1	1	Successful interactive anomaly detection
1	1	0	0	Stockbroker
1	1	0	1	Ordinary data mining
1	1	1	0	
1	1	1	1	Realistic ordinary science or student learning by help of an instructor in a lab

dark grey: no *learning* paradigms; white: most sceptic scenarios;

Online Machine Learning: Characterisation

Online machine learning is parsimonious when it comes to its input:
It processes it on-line.

And it can deal with adversarial scenarios.

It does not need intervention/experimentation.

It only needs supervision, i.e. feedback about the outcome.

Online Machine Learning: How it Works

How does it work? Basically, it is about a prediction tournament:

x_1	x_2	x_3	x_4	x_5	$\dots$
$pr_{1,1}$	$pr_{1,2}$	$pr_{1,3}$	$pr_{1,4}$	$pr_{1,5}$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$pr_{n,1}$	$pr_{n,2}$	$pr_{n,3}$	$pr_{n,4}$	$pr_{n,5}$	$\dots$

x_t are the true values; $pr_{1,t}, \dots pr_{n,t}$ are the predictions of different methods;

The accuracy of a prediction is measured via a loss function (within $[0, 1]$):

$$\ell(pr_{i,t}, x_t)$$

The performance of a method is measured via tracking its accuracy:

$$success_t(pr_i) = \frac{\sum_{1}^{t} (1 - \ell(pr_{i,t}, x_t))}{t}$$

Online learning algorithm: *success*-based weighting of pr_is

Online Machine Learning: Main Result

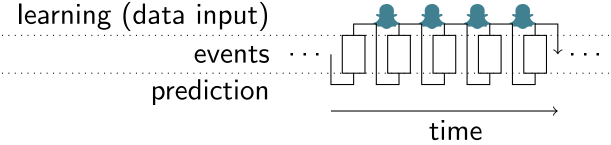
There is an important result about such an online learning algorithm:

Optimality of Online Learning

The algorithm is optimal, i.e. its *success*-rate is maximal in the limit.

This means that even if it is deceived (in favour of a competitor), it cannot be outperformed.

Its success in adversarial settings makes it fit also for spam detection: Spammers try to trick you into believing you did not receive spam from them.



Online Machine Learning: Application

What lurks behind emails/spam ...

... basically also lurks behind induction.

The theory of meta-induction of Schurz (2019) employs results of machine learning theory to overcome Hume's problem of induction.

Online Machine Learning: Application

So, AI can be also employed to approach epistemic problems.

This spans also over to the field of social epistemology.

Machine learning can be employed to address also the problems of:

- testimony
- peer disagreement
- judgement aggregation
- epistemic authority
- etc.

However, AI methods come not for free, they make assumptions.

Practically, they might pay back with success.

However, how can their assumptions accounted for theoretically or epistemologically?

A Social World for AI

Assumptions of Online Machine Learning

In order to achieve optimality, the discussed learning algorithms need to make the following assumptions: the tournament is about

- optimality vs. reliability
- long run vs. short run successes
- finite vs. infinitely many competitors
- accessible vs. non-accessible competitors
- continuous vs. discrete predictions
- bounded vs. unbounded losses
- convex vs. non-convex losses

(Almost) all of these assumptions have been discussed in the literature.

Some are really hard to come by (and bring in further problems).

Switching to a social setting $\Rightarrow$ makes assumptions natural

Example: Optimality

Hume's problem concerning the reliability of induction cannot be solved.

This is illustrated, e.g., by so-called *no free lunch theorems* (Wolpert 1996).

Switching to a social setting brings naturally competition/optimisation (vs. maximisation) with it.

It transforms the question about what is *best per se* to a question about what is *best in comparison with something*.

$\Rightarrow$

Example: Accessibility

Accessibility is a highly idealised assumption, particularly if one has real-world cases of prediction competitions in mind.

Switching to a social setting allows for de-idealisation.

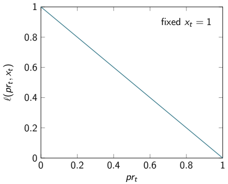
The idea is that the learning algorithm is not a competing predictor.

Rather, it is an aggregator, a predictor in the interest of all; it tries to get out the best for all through cashing out the current predictions of all.

$\Rightarrow$

Example: Convexity

Convexity of the loss function guarantees that the loss of a weighted average of some predictions does not exceed the weighted average of their losses.



Convexity is not only sufficient, but also necessary for optimality.
Sticking to the individual realm provides no argument for convex losses.
Switching to a social setting brings in such an argument: If we do not measure loss in a convex way, we have no guarantee for optimal aggregation.

 $\Rightarrow$ 

A Worry

Does this not simply amount to a proof-of-concept, showing that ML/AI can be relevantly used in a social epistemic setting?

That's definitely included. But it does not stop there.

I suggest a broader aim:
Not only think about how to best employ ML/AI in available social structures, but also devise social structures in order to best employ ML/AI.

Analogy:

- Democratic society: you can seek for infallible principles or simply safeguard societies against what we consider anti-democratic behaviour
- Justification of AI: you can seek for infallible principles or simply safeguard against what we consider epistemic failures

Summary

- AI plays an increasingly important role for society.
- One particular form of AI, online machine learning, has also important epistemic impact.
- But it is based on quite idealised assumptions.
- In a social setting, these assumptions can be de-idealised.
- Sometimes this asks not only for devising new forms of AI, but also for devising new forms of social structure.

References I

CERN (2014). *Higgs Boson Machine-Learning Challenge*. URL: https://home.cern/news/news/computing/higgs-boson-machine-learning-challenge (visited on 2014-11-19).

Chui, Michael, Harrysson, Martin, Manyika, James, Roberts, Roger, Chung, Rita, Nel, Pieter, and van Heteren, Ashley (2018). "Applying Artificial Intelligence for Social Good". In: *McKinsey Global Institute Notes from the AI Frontier* November 28. URL: https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good#.

Hauser, Larry (2012). "Artificial Intelligence". In: *Internet Encyclopedia of Philosophy*. Ed. by Fieser, James and Dowden, Bradley.

Kleinberg, Jon, Ludwig, Jens, and Mullainathan, Sendhil (2016). "A Guide to Solving Social Problems with Machine Learning". In: *Harvard Business Review* December 08. URL: https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning#.

Russell, Stuart and Norvig, Peter (2020). *Artificial Intelligence: A Modern Approach*. Fourth Edition. Boston.

Schurz, Gerhard (2019). *Hume's Problem Solved. The Optimality of Meta-Induction*. Cambridge, Massachusetts: The MIT Press.

Shalev-Shwartz, Shai and Ben-David, Shai (2014). *Understanding Machine Learning. From Theory to Algorithms*. Cambridge: Cambridge University Press.

Wolpert, David H. (1996). "The Lack of A Priori Distinctions Between Learning Algorithms". In: *Neural Computation* 8.7, pp. 1341–1390. DOI: 10.1162/neco.1996.8.7.1341.